CPAD
California Protected Areas Database

CPAD PROJECT WORKING PAPER

DATA QUALITY IN CPAD – REVIEW AND ISSUES

Prepared by

Larry Orman, Project Director
and John Kelly, Sr. GIS Specialist

May 15, 2013

CONTENTS

GreenInfo Network

www.greeninfo.org

SUMMARY

The California Protected Areas Database (CPAD) is a statewide inventory of parks and other protected open lands owned by agencies and organizations for the purpose of maintaining these open space uses (see www.CALands.org for more information). In late 2012, GreenInfo Network, the originator of CPAD, began a two year project to improve this data set, with funding support from the California Strategic Growth Council through the USGS Gap Analysis Program. This report is one of a series being issued through this project.

The focus of this report is on quality – for the data in CPAD, how good is it? For the four areas we assessed, the following summarizes our findings:

Completeness – CPAD appears to capture well over 95% of all protected land acreage in California and may well have captured in excess of 99%. As far as number of parks/preserves goes, CPAD appears to have well over 80%, possibly 95% or more (this measure is good for assessing city parks which are numerous but small in total acres).

Spatial accuracy – This is defined as the degree to which CPAD is aligned to parcels (or any other chosen base layer). Overall, CPAD holdings average a distance from nearest parcel of 10 meters – however, this figure is skewed by a few large holdings (national forests) that, once issues with parcel alignment policy are resolved for them, should bring the total down to five or fewer meters. It appears that CPAD is closely aligned to parcels in most of the state, except for: far northern Calif. national forests; the central Sierra national forests, and the deserts (BLM areas). There are a dozen or so clusters of parcel mis-alignments around several reservoirs and wetlands reserves.

GreenInfo developed an innovative analytical tool to help measure accuracy by calculating cumulative distance from holding vertices to nearest parcels. We expect to run this again in fall 2013 and should see major improvements due to editing work based on these measures.

Attribute completion – With the exception of defining the alignment source for holdings, all CPAD fields that require values are 95%+ completed, with most being 100% completed. A number of fields have much lower completion rates, but these are either fields where only a smaller number of completes is expected (e.g., managing agency ID, as only a relatively few holdings are managed by other than their owner), or in a couple of cases, where the field is a legacy field not actively used ("primary purpose") or has not yet been fully populated (date of acquisition, which is not part of this project).

Errors - This covers wrongly included holdings and digitizing errors. We do not have an effective means to globally assess errors, but have focused on reviewing older legacy data which is prone to more errors than modern data. We also have clarified our internal editing policies to guide decisions about including/revising data.

I. INTRODUCTION

The quality of protected areas data matters to users. This memo summarizes the state of the California Protected Areas Database (CPAD) and outlines issues and directions for improvements in both data and measuring processes.

We have included completeness, spatial accuracy, attribute completion and errors as the major topics for this memo, with a major focus on spatial accuracy.

The importance of spatial accuracy is largely a function of the geographic scale at which users expect reliable results from use of a GIS data set. A review of the proximity of protected lands to river corridors throughout California ("how many acres are within one mile of Wild and Scenic riverways?") might tolerate data that was not precisely located as to geographic features and/or land ownership. On the other hand, an assessment of the number of park acres within a quarter mile of any point in an urban neighborhood might need data with a small margin of inaccuracy.

However, spatial accuracy is also a function of budgets available for data creation and refinement – even with a demanding need for accuracy, funding may not exist for the data work required to meet that need.

Our analysis and findings on spatial accuracy and the other factors follow.

II. GIS DATA ACCURACY

Geospatial data accuracy is generally approached by comparing real world values to data created in GIS. There are usually three elements taken into account: positional and attribute accuracy, data quality (lineage, completeness, logical consistency) and error.

Evaluating data accuracy generally is a also function of geographic scale and budget - data that has a strong use case for being extremely accurate may require a much greater budget for fine grained editing than data that is primarily used at small scales.

In the case of CPAD, a statewide data collection of many thousands of geographic objects which is not the primary land record system for any agency and which has a limited ongoing budget target (approximately $100,000/year required to maintain data), GreenInfo believes the requirements for accuracy should be focused on these four elements:

1. **Completeness of the collection** – do we have close to all relevant parks and open spaces represented geographically?
2. **Relation of geometry to underlying base map** – assuming CPAD is mapped primarily to assessor parcels, how close do its holdings' boundaries match parcel lines? If not, is there a reasonable alternative base geometry (PLSS, surveys, etc.) and how does CPAD match that base?
3. **Quality and consistency of attributes** – are data attributes correct (unit name, agency owner/manager, etc.), complete and are they implemented consistently throughout the data set?
4. **Errors** - are parcels or owners included in the data who should not be in it; are holdings wrongly digitized or transferred in; and were CPAD data policies fairly and consistently applied to defining the area of holdings or their attributes?

We note at the outset of this assessment that CPAD data are not original source data – parks and open spaces are defined in CPAD based on how they exist in other data sets and are often spatially adjusted into CPAD. Therefore, measures of spatial accuracy that might be appropriate for original data collection are not relevant here.

Though discussed in a separate memo, it is also important to recognize that CPAD data are intended to be aligned spatially to underlying boundaries of property ownership (assessor parcels). While there are some exceptions to this approach and issues with assessor parcel data (especially in rural areas), in general the measure of spatial accuracy in CPAD is alignment with parcel boundaries.

III. EVALUATING DATA ACCURACY IN CPAD

ACCURACY METRICS – COMPLETENESS OF CPAD

CPAD contains holdings for almost 1,000 agencies and organizations, and our ability to understand whether we have all holdings of these agencies is limited – in many cases with smaller agencies, we don't know who to contact, don't have resources for individual outreach to all these agencies and, on top of this, many of these agencies don't have GIS data or even authoritative published park/open space maps available.

However, since only a tenth of CPAD agencies own 99 percent of acres of holdings we can assess whether we have reasonable current data for these agencies in CPAD. For the other 880 agencies (which have 65% of the units in CPAD, mostly smaller urban parks), we have to make best estimate assumptions, as we often do not have source data for these. Finally, we have not yet addressed how we will track updates once we "catch up" to what should be in CPAD as of a current year.

RESULTS: As of early 2013, GreenInfo believes it has very complete data in CPAD for the top 100 agencies that hold 99 percent of acres. Within the top 25 agencies (90% of acres), we have been more active in ensuring we have updated data. For all the smaller agencies, we know that we do not always have complete data (and there may be alignment issues in some cases), but we have sampled reviews of cities in particular and have generally found that we have almost all such data in most areas (San Diego, Orange and Riverside counties were exceptions but are being addressed in 2013).

ACCURACY METRICS – PARCEL ALIGNMENT

Evaluating the data accuracy elements noted above against the actual CPAD data requires identifying cost-effective, valid and useful methods. GreenInfo is applying the following three strategies – qualitative review (visual), quantitative analysis and random visual checking:

1. QUALITATIVE evaluation of fit to county parcel geometry – we have reviewed each of California's 58 counties to visually judge the degree of fit between CPAD and parcels. We use a High, Medium, Low ranking to indicate fit with parcels. This ranking does not yet account for whether there is major effort required to fix mis-alignments (e.g. Riverside County has a very large and complex inventory of protected areas and is generally not yet aligned; Kern County has a very small inventory with just a few apparent alignment issues).

RESULTS: We conducted this sort of visual review of each county and concluded that almost half of counties were quite consistent (over 90% of CPAD holdings) with parcel geometry, another quarter were reasonably consistent (75%+ of CPAD holdings) and a quarter of counties had significant inconsistencies (mostly Sierra and Desert counties where USFS and BLM parcel alignments have issues (or have not been so aligned).

After conducting our next measure, we have concluded that visual assessment should be a check against our quantitative measure, rather than its own measure.

2. QUANTITATIVE Measures of Individual Holding Alignment to Parcels - GreenInfo developed a unique method using spatial statistics to characterize the fit to parcel lines of each holding in CPAD, which is also aggregated to county and statewide levels. In general, our algorithm measures the distance between every vertex in a CPAD holding and the closest point of the nearest parcel geometry. This measure establishes an average of these distances which becomes a score that is attached to each CPAD holding. For counties, the average or other measure of all CPAD holdings can be created to define the general state of any particular county. But most important, this method highlights CPAD holdings that have a poor degree of fit to parcels, allowing us to quickly pinpoint where editing for parcel alignment should be focused.

A summary of the average distance to parcel lines in each county is provided in the appendix. It ranged from close to 0 to up to 50 meters.

More telling is that CPAD data can be displayed in GIS with these alignment rankings, creating a powerful visual tool for rapid accuracy assessment, in relation to parcel data.

Method Details: GreenInfo computed the distance from each vertex of each holding to the nearest alignment line (the shortest distance to a parcel line, waterbody, or coastline). Here are the steps we took to create this assessment (details are in the appendix):

1. Convert CPAD polygons (holdings) to vertex points
2. Convert CPAD polygons to lines and buffer these lines so that they can be intersected with 58 county parcel datasets to "pre-select" only the parcels that we use in the analysis.
3. Convert parcels intersecting CPAD to lines (this avoids distance=0 for vertices lying within a parcel, instead giving us the distance to the boundary of the parcel).
4. Combine the parcel lines with waterbody lines and coastline.
5. For each vertex, compute the distance to the nearest alignment line (roughly 4 million vertices; 50 hours of computer processing).
6. Summarize average of the distances for each holding and join this value to CPAD holdings.

RESULTS: In this first run of our method, we found that the statewide average of the distance of all holdings to the nearest parcel is 10 meters. The range by county is near 0 to upwards of 50 meters. However, we believe that overall, CPAD is actually much closer to parcels due to:

In counties where there are many lakes (esp. the Sierra), we get abnormally high variances – this should be corrected in future runs

In some coastal counties with many offshore islands, these islands are at variance with parcel geometry and drag down overall terrestrial accuracy

For some large holdings (e.g., national forest lands in CPAD which do not usually have interior parcels), small variances from source data can lead to abnormally large parcel variance scores (this was particularly true in Monterey County, where a few irregular USFS lines cause a very large CPAD forest holding to be shown as low accuracy, when almost all other geometry was perfectly aligned to parcels.

Nonetheless, there are also smaller clusters of parcel alignment issues that are real and will need editing (there are some Central Valley wetland reserve holdings and reservoirs at issue, along with BLM and USFS lands generally) especially in the Sierra and north state, as well as the desert – as well as scattered units that need adjustment.

Once the major high variance parcels are corrected, a rerun of this analysis should be done, to create a longer-term framework for guiding editing. With present resources, this analysis is requires significant resources and is unlikely to be able to be run more than annually (CPAD is currently updated twice yearly).

Random visual check of holdings - The final method, GreenInfo has also tested an approach that selects random holdings in each of three size ranges and visually inspects

the difference between the parcel geometry and the CPAD holding. We conducted this inspection on 100 holdings in each size range.

Results: We found this to be inferior to the above two methods, as the possibilities of variation in each holding are quite large and we had a hard time determining how to generalize from these experiences. In light of the parcel distance method, we conclude that random checking is not likely to be effective.

ACCURACY METRICS – Attribute Completion

Another important measure of CPAD accuracy is the degree to which each of its attribute fields are completed. The following table summarizes CPAD 1.9 (most recent release) and shows that, by and large, CPAD field completion is very high.

FIELD	# Holdings	Complete Percent	NOTE
AGNCY_NAME	55,583	100%	
UNIT_NAME	55,583	100%	
ACCESS_TYP	55,583	100%	
GIS_ACRES	55,583	100%	
COUNTY	55,583	100%	
AGNCY_LEV	55,583	100%	
AGNCY_WEB	54,962	99%	
SITE_WEB	13,892	25%	Low priority, does not exist for many sites
LAYER	55,583	100%	
MNG_AGNCY	1,058	2%	Only few sites have sep. mgmt agcy
LABEL_NAME	55,472	100%	
OWN_TYPE	55,583	100%	
SITE_NAME	55,583	100%	
ALT_SITE_N	157	0%	Few alternative site names
LAND_WATER	55,583	100%	
SPEC_USE	2,263	4%	Few special uses
HOLD_NOTES	10,472	19%	Used only when needed
CITY	23,206	42%	Mainly for city holdings
DESG_AGNCY	51,114	92%	
DESG_NAT	54,490	98%	
PRIM_PURP	33,575	60%	Older legacy field, not active now
APN	2,531	5%	Not an active data field
HOLDING_ID	55,583	100%	
UNIT_ID	55,583	100%	
SUPERUNIT	55,583	100%	
AGNCY_ID	55,583	100%	
MNG_AG_ID	1,054	2%	See "MNG_AGNCY"
AL_AV_PARC	51,825	93%	
DATE_REVIS	55,583	100%	
SRC_ALIGN	45,778	82%	Improve coverage in this field
SRC_ATTR	5,800	10%	Only limited use - source of attribute data
D_ACQ_YR	2,654	5%	Acquisition dates only for a few sites

ACCURACY METRICS – Errors

As noted at the beginning of this memo, there are generally three types of errors that might occur in CPAD data:

- Erroneous inclusion of holdings
- Inaccurate spatial and attribute data for included holdings
- Ineffective application of CPAD editing policies

We know that some of our legacy data (particularly before 2008) is erroneous – many of these holdings were quickly defined into CPAD due to meager resources, and some were digitized from USGS quads or other base maps because parcel and air imagery data was not available. We have extracted much of this (its source data was noted in our records) and subject it to review, removing many questionable holdings and refining others to parcel lines.

Alignment issues are discussed above.

We certainly have some attributes that are lacking – the most significant of these would include who owns a holding, and what its public access is. Our general sense is that at this point, there are few errors in regard to owning agencies, although we have a number of holdings for which access and other attributes might be incomplete. General state of attributes is covered above.

Finally, we have reviewed, better documented and improved our editing policies. However, it is not possible to do detailed quality assurance on every CPAD holding, given our available resources. We have regular meetings with editing staff and have reasonable training procedures. There is no question that there is room for improvement here, but on balance, we have found that there is a balance that is needed between an overly detailed focus on fine grain data and developing other aspects of CPAD (educating others about its availability, building relationships with key users and contributing agencies, etc.).

RESULTS: CPAD is not error-free by any means, but great progress has been made in correcting past issues with legacy data and appropriate procedures guide us in our editing work.

I. APPENDIX – Methods for Parcel Accuracy Assessment

As noted above, the following are the general steps we took in developing a method for analyzing parcel alignment of CPAD holdings:

1. Convert CPAD polygons (holdings) to vertex points
2. Convert CPAD polygons to lines and buffer these lines so that they can be intersected with 58 county parcel datasets to "pre-select" only the parcels that we use in the analysis.
3. Convert parcels intersecting CPAD to lines (this avoids distance=0 for vertices lying within a parcel, instead giving us the distance to the boundary of the parcel).
4. Combine the parcel lines with waterbody lines and coastline.
5. For each vertex, compute the distance to the nearest alignment line (roughly 4 million vertices; 50 hours of computer processing).
6. Summarize average of the distances for each holding and join this value to CPAD holdings.

Details of steps (from automated scripting created):

1. Convert cpad polygons (holdings or units) to points (for each vertex): cpad_verts

```
CREATE TABLE cpad15_verts as
        SELECT holding_id, (ST_DumpPoints(geom)).*
        FROM cpad15_2010_06
SELECT 3928374
```

Time: ~1min

Create "unique ids":

```
CREATE SEQUENCE cpad15_verts_gidjk_seq;
ALTER TABLE cpad15_verts ADD COLUMN gid_jk int NOT NULL DEFAULT
nextval('cpad15_verts_gidjk_seq');
```

2. Convert cpad polygons to lines... to be buffered by search radius... used in next step...

```
CREATE TABLE cpad15_lines_buff200m
  AS SELECT
    holding_id,
    ST_Buffer(ST_Boundary(geom),200) as geom
  FROM cpad15_2010_06;
```

SELECT 51982
Time: ~13min
Export to shp.

3. ... to intersect with 58 county parcel datasets to "pre-select" only the parcels that we used in the analysis: parcels_intersecting_cpad.shp [Currently using an arcpy script; takes about 6 hours.]

```
C:\Python27\ArcGISx6410.1\python.exe
U:\Users\JK\U_JK_parcels\statewide\selectAllParcelLayersIntersectingCPAD_copy.py
import arcpy
mxd = arcpy.mapping.MapDocument(r'U:\Users\JK\parcels_state.mxd')
df = arcpy.mapping.ListDataFrames(mxd)[0]
lyrs = arcpy.mapping.ListLayers(mxd)
cpad_lyr = arcpy.mapping.ListLayers(mxd,"cpad",df)[0]

for lyr in lyrs:
  print lyr.name
  if lyr.name != "cpad": #cpad =
P:\proj_a_d\CPAD\CPAD_project\data\temp\SpatialAssessment\CPAD_Holdings_lines_
Buffer100m_ginserver.shp
    arcpy.SelectLayerByLocation_management(lyr, "INTERSECT", cpad_lyr)

     #create this empty shp beforehand (srs 3310), with no fields (rather, a single dummy
field)
     arcpy.Append_management
(lyr,'U:\Users\JK\U_JK_parcels\statewide\parcels_intersecting_cpad15_2010.shp',"NO_
TEST")
################ end python script

ogr2ogr -f "PostgreSQL" -t_srs "EPSG:3310" PG:"host=localhost user=postgres
dbname=cpad_verts password=ginfo116" parcels_intersecting_cpad15_2010.shp -nln
parcels3ogr_cpad15 -overwrite -nlt multipolygon
```

4. Convert parcels intersecting cpad to lines. (We don't want distance=0 for vertices lying within a parcel--we want the distance to the boundary of the parcel.) These must be converted to simple polygons first (from multipolygons):

```
CREATE TABLE parcels3ogr_lines
  AS SELECT
    ST_Boundary(wkb_geometry) AS geom
```

```
  FROM parcels3ogr_cpad15;

SELECT 3032250
Time:  ~2min15sec

ALTER TABLE parcels3ogr_lines ADD COLUMN class varchar(30);
UPDATE parcels3ogr_lines SET class = 'parcels';
```

5. Combine the parcel lines with waterbody lines and coastline.

```
INSERT INTO parcels3ogr_lines (class, geom) SELECT class, st_boundary from
waterbody_lines;

CREATE INDEX idx_parcels3ogr_lines ON parcels3ogr_lines USING GIST ( geom );
```

6. For each vertex, compute the distance to the nearest alignment line. (Roughly 4 million vertices)

```
CREATE TABLE vert_dist AS
  SELECT DISTINCT ON(g1.gid)  g1.gid AS vert_id, g1.holding_id AS holding_id, g2.gid AS
union_id, g2.class AS class,
  ST_Distance(g1.geom,g2.st_boundary) AS dist_m
  FROM cpad_verts AS g1, parcels_lines AS g2
  WHERE ST_DWithin(g1.geom,g2.st_boundary, 1000)
  ORDER BY g1.gid, ST_Distance(g1.geom,g2.st_boundary);
SELECT 4241200
Time: ~50hrs.

CREATE TABLE vert_dist15 AS
  SELECT DISTINCT ON(g1.gid_jk)  g1.gid_jk AS vert_gid_jk, g1.holding_id AS
holding_id,  g2.class AS class,
  ST_Distance(g1.geom,g2.geom) AS dist_m
  FROM cpad15_verts AS g1, parcels3ogr_lines AS g2
  WHERE ST_DWithin(g1.geom,g2.geom, 200)
  ORDER BY g1.gid_jk, ST_Distance(g1.geom,g2.geom)
  ;
SELECT 3805294
Time:  ~2h15m
```

7. Summarize the near distances for each holding; join to CPAD holdings.

```
CREATE TABLE avg_dist_holding_no_water AS
SELECT holding_id, avg(dist_m), min(dist_m), max(dist_m)
FROM vert_dist
WHERE class='parcels'
GROUP BY holding_id;
```

Time: 30 seconds

8. Test/Visualize Vertex Scores

```
ALTER TABLE cpad15_verts ADD COLUMN class varchar(30);
ALTER TABLE cpad15_verts ADD COLUMN dist_m double;
```

Create Indexes!

```
CREATE INDEX cpad15_verts_gid_jk ON cpad15_verts (gid_jk);
CREATE INDEX cpad15_verts_gid_jk ON cpad15_verts (gid_jk);
CREATE INDEX vert_dist15_vert_gid_jk ON vert_dist15 (vert_gid_jk);
```

Join dist results to vert pts:

```
UPDATE cpad15_verts AS cv
SET dist_m = vd.dist_m, class = vd.class
FROM vert_dist15 AS vd
WHERE cv.gid_jk = vd.vert_gid_jk
```

Export some to shp:

```
ogr2ogr -f "ESRI Shapefile" C:\temp\cpad15_verts_test.shp PG:"host=localhost
user=postgres dbname=cpad_verts password=ginfo116" -sql "SELECT * from
cpad15_verts where holding_id<100"
```